ECON 5033 Econometrics I – Lecture 3

Multiple Linear Regression Model – Part I

Eric S. Lin

Department of Economics, National Tsing Hua University

October 8, 2024

Simple Linear Regression Model

- In the previous lecture we covered the simple linear regression model, i.e., only one regressor (X_i)

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- Now consider the matrix expression:

$$Y_{n\times 1} = \iota_{n\times 1}\alpha_{1\times 1} + X_{n\times 1}\beta_{1\times 1} + \varepsilon_{n\times 1}$$

- How about $Y_i = \alpha + \varepsilon_i$?
- Now, let's extend the model to allow for numerous regressors.
- Computation by hand will be infeasible.
- Expression using summation will be tedious. [will see 2×2 case with $X = (\iota : X_2 : X_3)$ in deviation form]

Multiple Linear Regression Model

- Consider the following setup:

$$Y_i = X_i'\beta + \varepsilon_i,$$

  - Y_i and ε_i are scalars, β is a $k \times 1$ vector and X_i is a $k \times 1$ vector.

- Note that $X_i' = (1 : X_{2i} : ... : X_{ki})$
- In matrix form, we have:

$$Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \varepsilon_{n\times 1},$$

$$X_{n\times k} = \begin{bmatrix} X_{11} & X_{21} & \dots & X_{k1} \\ \dots & \dots & & \dots \\ \dots & \dots & & \dots \\ X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & \dots & X_{k1} \\ \dots & \dots & & \dots \\ \dots & \dots & & \dots \\ 1 & X_{2n} & \dots & X_{kn} \end{bmatrix}$$

Classical Assumptions for Multiple Regression Models

Similar to the Assumptions in the simple linear regression model:

1. $\mathrm{E}[Y|X] = X\beta$
2. $\mathrm{E}[\varepsilon|X] = 0$
3. $\mathrm{E}[\varepsilon\varepsilon'|X] = \sigma^2 I_n$
4. $\mathrm{E}[\varepsilon_i\varepsilon_j|X] = 0$ for all $i \neq j$.
5. The data matrix X is of full column rank.

- Assumption 5 is also known as the identification condition..
- It is equivalent to say $\text{rank}(X) = \rho(X) = k$, or no multicollinearity.

Identification

Example

Consider the demand for cars.

$$\text{Exp}_i = \alpha + \beta_1 \text{sex}_i + \beta_2 \text{H\_Inc}_i + \beta_3 \text{W\_Inc}_i + \beta_4 \text{F\_Inc}_i + \beta_5 \text{age}_i + \varepsilon_i.$$

Example

Consider the effect of admission methods.

$$\text{GPA}_i = \alpha + \beta_1 \text{gender}_i + \sum_{j=1}^{5} \gamma_j \text{SAT}_{ji} + \tau \text{TSAT}_i + \delta \text{area}_i + \varepsilon_i.$$

Identification: $k = 3$

Example

Suppose $X_{2i} = 2X_{3i}$ for all i. Then

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \\ &= \beta_1 + \beta_2 2X_{3i} + \beta_3 X_{3i} + \varepsilon_i \\ &= \beta_1 + \beta_3^* X_{3i} + \varepsilon_i, \quad \text{Here } \beta_3^* = 2\beta_2 + \beta_3, \end{aligned}$$

where β_2 and β_3 cannot be separately identified. We can only estimate the coefficient up to β_3^*, that is $\hat{\beta}_3^*$.

Identification for the Case of $k = 3$

Example

Let's re-consider the case of $k = 3$.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i.$$

One could show

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{3i}^2\right) - \left(\sum y_i x_{3i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{3i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2}$$

$$\hat{\beta}_3 = \frac{\left(\sum y_i x_{3i}\right)\left(\sum x_{2i}^2\right) - \left(\sum y_i x_{2i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{3i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2},$$

where the lowercase letters denote deviations from sample mean values.

Estimation

- Given the quadratic loss function, the OLS estimator of β is

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}(Y - X\beta)'(Y - X\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2.$$

- The objective function is

$$\varepsilon'\varepsilon_{(1\times 1)} = Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta. \tag{1}$$

- FOC and SOC of (1) are

$$\frac{\partial\varepsilon'\varepsilon}{\partial\beta} = -2X'Y + 2X'X\beta = 0$$

$$\frac{\partial\varepsilon'\varepsilon}{\partial\beta\partial\beta'} = 2X'X \quad (\text{positive definite})$$

- The minimizer is given by:

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y.$$

Some Comments

- OLS estimator is also the Method of Moment (MoM) estimator.

$$\mathrm{E}[X_i\varepsilon_i] = 0 \iff \mathrm{E}[X_i(Y_i - X_i'\beta)] = 0$$

- If we consider another loss function, say $|\cdot|$, then

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left| Y_i - X_i'\beta \right| .$$

  $\tilde{\beta}$ is called a Least Absolute Deviations (LAD) estimator. In this case, zero median for error term ε_i is assumed.
- LAD is more robust (to outliers) than OLS

Some Comments

- Look at the familiar normal equation (i.e., FOC),

$$X'Y - X'X\hat{\beta} = 0 \Leftrightarrow X'(Y - X\hat{\beta}) = 0 \Leftrightarrow X'e = 0.$$

- Note that we call $X\hat{\beta} = \hat{Y}$ the predictive value or fitted value of Y.
- The residual vector e is defined as the difference between the observed value Y and the fitted value $\hat{Y}$.
- Estimate vs. estimator

Geometric Interpretation

Definition

Addition, multiplication (stretch, shrink, and reverse) of vectors.

Definition (Graphical presentation)

The orthogonal complement of $\delta(X)$ in Euclidean space E^n, which is denoted $\delta^{\perp}(X)$, is the set of all vectors w in E^n that are orthogonal to everything in $\delta(X)$. We use $\delta(X)$ to represent the subspace associated with the k columns of X. Formally,

$$\delta^{\perp}(X) \equiv \{w \in E^n | w'z = 0 \text{ for all } z \in \delta(X)\}.$$

Geometric Interpretation

Example

Linear regression model: $Y = X\beta + \varepsilon$.

Example

Graphical illustration of orthogonality between residuals and fitted values.

Example

Assume $\delta(X) = \delta(X_1, X_2)$, graph the projection of Y on two regressors.

Example

Breakdown of TSS into ESS and RSS by Pythagoras' Theorem. That's the uncentered R^2.

Geometric Interpretation

Example

Consider the following X matrix, which is of 5×3 :

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

One can see that $x_1 = ax_2 + bx_3$. So that

$$\delta(X) = \delta(x_1, x_2) = \delta(x_1, x_3) = \delta(x_2, x_3).$$

Projection Matrix & Residual Making Matrix

- Let's define the following projection matrix:

$$P_X = X\left(X'X\right)^{-1}X',$$

- Projecting the observed Y value onto the space of the columns of independent variables X produces the set of the fitted values.

$$P_X Y = X\left(X'X\right)^{-1}X'Y = X\hat{\beta} = \hat{Y}.$$

- Projecting the observed Y value onto complementary space of the columns of independent variables X produces the OLS residual vector.

$$M_X Y = (I - P_X)\,Y = Y - X\left(X'X\right)^{-1}X'Y = Y - \hat{Y} = e,$$

where M_X is known as the residual making matrix.

Projection Matrix & Residual Making Matrix

- Any vector already within the span of X will be projected into itself [$s = Xb$ for some b]: $P_X s = s$
- Any vector (say w) in the subspace orthogonal to that spanned by X, $w \in \delta^{\perp}(X)$: $P_X w = 0$
- $\delta^{\perp}(X)$ coincides with the image of M_X
  1. $\delta^{\perp}(X)$ is contained in the image of M_X : $M_X w = w$
  2. All vectors in the image of M_X belong to $\delta^{\perp}(X)$: $(M_X Y)'X = 0$
- What is analytically convenient need not be computationally useful, e.g., formation of fitted values; dimension of P

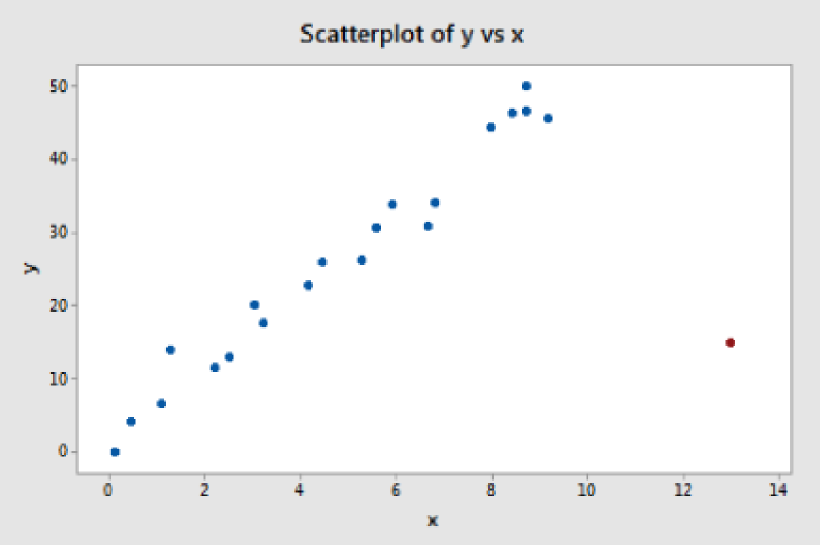
Properties of the Projection Matrix

- Note that $\hat{Y}$ is the orthogonal projection of Y onto $\delta(X)$ and e is the orthogonal projection of Y onto $\delta^{\perp}(X)$.
- P_X and M_X are sometimes called complementary projections. $[P_X Y + M_X Y = Y]$
- It is easy to see the following properties. [Verify!]
  1. $X'e = 0$, $\hat{Y}'e = 0$.
  2. $P_X X = X$, $M_X X = 0$, $P_X M_X = 0$ (annihilate each other).
  3. Idempotent property: $P_X P_X = P_X$, $M_X M_X = M_X$.

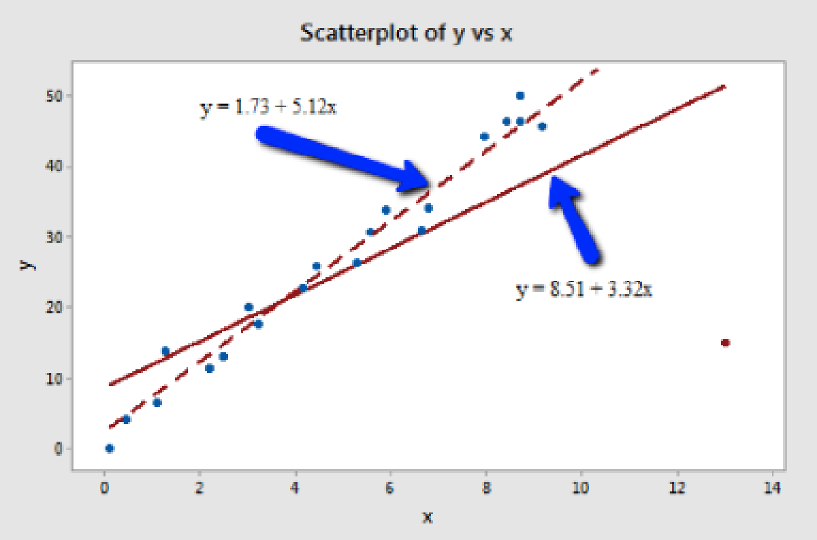
Influential Observation

- Recall the projection matrix P_X
- Leverage of observation i: $[P_X]_{ii} = \iota_i' P_X \iota_i \equiv h_i$
- Note that ι_i is a zero vector with "1" in the i^{th} row
- We may use h_i to measure an outlier or a high leverage point

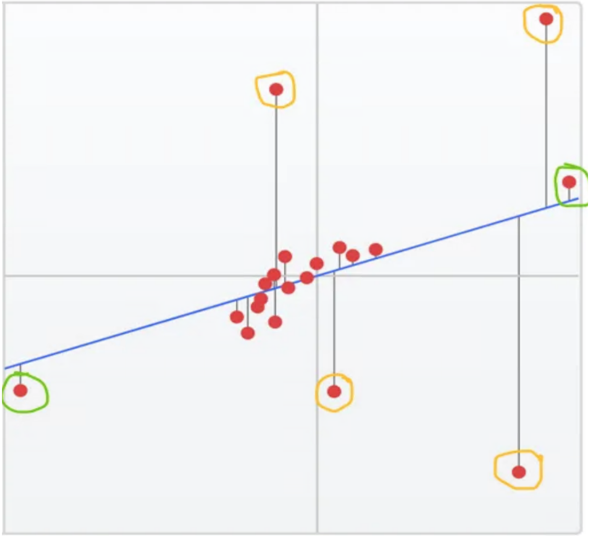
Influential Observation – Scatter Plot



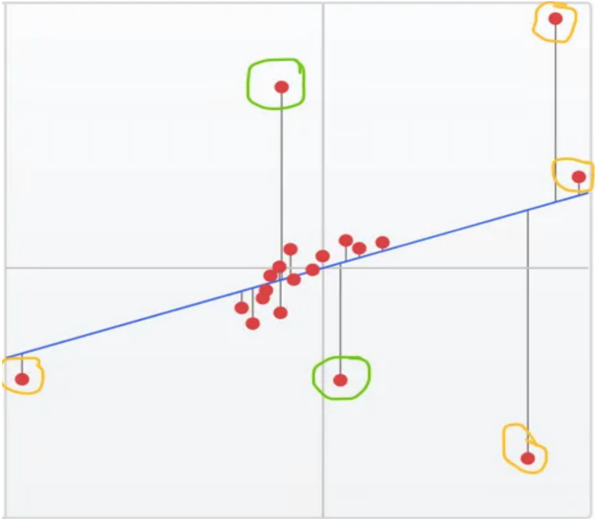
Influential Observation – Regression Result



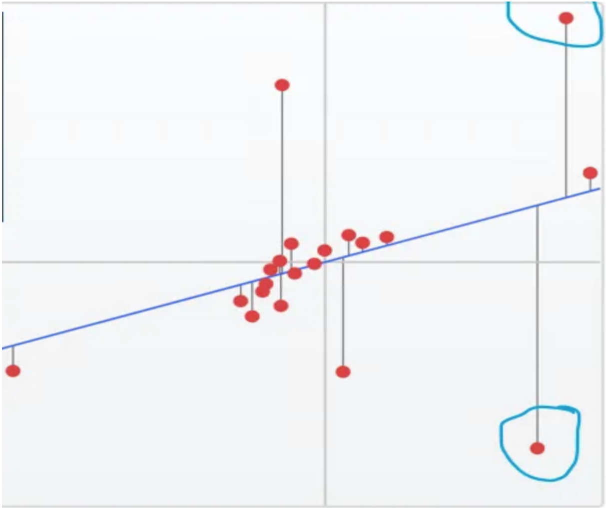
Outliers



Leverage Points



Influential Points



Frisch-Waugh-Lovell

- Let's assume the following specification.

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon, \tag{2}$$

  where X_1 and X_2 are two groups of regressors, i.e., $X = (X_1 : X_2)$. Assume that X_1 and X_2 are of full ranks k_1 and k_2 respectively.
- In some case, one would like to know either β_1 or β_2 only but not both.
- Deviation from mean regression is a good example.

Frisch-Waugh-Lovell

Theorem (FWL)

The OLS estimators in (2) are equivalent to the followings.

$$\hat{\beta}_1 = \left[X_1' M_2 X_1\right]^{-1} X_1' M_2 Y$$
$$\hat{\beta}_2 = \left[X_2' M_1 X_2\right]^{-1} X_2' M_1 Y,$$

where $M_1 = I - X_1 \left(X_1' X_1\right)^{-1} X_1'$ *and* $M_2 = I - X_2 \left(X_2' X_2\right)^{-1} X_2'$.

Frisch-Waugh-Lovell

Proof.

We prove the coefficient of β_2. Note that Y could be decomposed as

$$Y = P_X Y + M_X Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + M_X Y. \tag{3}$$

Premultiplying (3) by $X_2' M_1$ gives

$$X_2' M_1 Y = X_2' M_1 X_2 \hat{\beta}_2.$$

Here we use the fact that $X_2' M_1 X_1 \hat{\beta}_1 = 0$ and $X_2' M_1 M_X Y = 0$. (why?) The result follows. Another proof? □

Frisch-Waugh-Lovell: Direct Proof

Proof.

$$
\begin{aligned}
Y &= X_1\beta_1 + X_2\beta_2 + \varepsilon = [X_1 : X_2][\beta_1' : \beta_2']' \\
\hat{\beta} &= [\hat{\beta}_1' : \hat{\beta}_2'] = [[X_1 : X_2]'[X_1 : X_2]]^{-1}[X_1 : X_2]'Y \\
&= \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} [X_1 : X_2]'Y
\end{aligned}
$$

$\square$

Application of FWL

Example

FWL theorem tells us that $\hat{\beta}_2$ could be computed from regressing $M_1 Y$ on $M_1 X_2$, which means regressing the residuals of Y on X_1 on the residuals of X_2 on X_1, i.e.,

$$M_1 Y = M_1 X_2 \beta_2 + M_1 \varepsilon.$$

Example

Deseasonalization. Consider the model with seasonal component.

$$\begin{aligned} Y &= \alpha_1 s_1 + \alpha_2 s_2 + \alpha_3 s_3 + \alpha_4 s_4 + X_2 \beta_2 + \varepsilon \\ &= X_1 \beta_1 + X_2 \beta_2 + \varepsilon. \end{aligned}$$

According to FWL theorem, $M_1 Y$ is a form of seasonal adjustment or de-seasonalization.

Application of FWL

Example

Consider the two groups regressor model in (2) and let $col(X_2) = 1$. One would get

$$\hat{\beta}_2 = \frac{X_2' M_1 Y}{X_2' M_1 X_2} = \frac{X_2^{*\prime} Y^*}{X_2^{*\prime} X_2^*} = \frac{X_2^{*\prime} Y^*}{\sqrt{X_2^{*\prime} X_2^*}\sqrt{Y^{*\prime} Y^*}} \frac{\sqrt{Y^{*\prime} Y^*}}{\sqrt{X_2^{*\prime} X_2^*}} = r_{2,Y}^* \frac{\hat{\sigma}_{Y^*}}{\hat{\sigma}_{X_2^*}},$$

where $r_{2,Y}^*$ is the partial correlation coefficient between X_2 and Y conditional on X_1. Remember in the simple linear regression model $(\mathrm{col}(X_1) = \mathrm{col}(X_2) = 1)$,

$$\hat{\beta} = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2} = \frac{\hat{\sigma}_{x,y}}{\hat{\sigma}_x^2} = r_{x,y} \frac{\hat{\sigma}_y}{\hat{\sigma}_x}.$$

Long and Short Regressions

- According to Goldberger (1991), the long and short regressions are defined accordingly:

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \quad [\text{long regression}] \\ Y &= X_1\gamma_1 + v \quad [\text{short regression}] \end{aligned}$$

- Typically, $\beta_1 \neq \gamma_1$, except in special cases. To see this

$$\begin{aligned} \gamma_1 &= E[X_1'X_1]^{-1}E[X_1'Y] \\ &= E[X_1'X_1]^{-1}E[X_1'\{X_1\beta_1 + X_2\beta_2 + \varepsilon\}] \\ &= \beta_1 + \Gamma_{12}\beta_2, \end{aligned}$$

  where Γ_{12} is the coefficient matrix from a projection of X_2 on X_1.
- Thus the short and long regressions have different coefficients on X_1. They are the same only under one of two conditions.
- Sometimes, $\Gamma_{12}\beta_2$ is know as omitted variable bias (OVB).

R-square revisited

- Recall the r^2 in simple linear regression model
- In multiple regression model, we use R^2 to denote the goodness of fit.

$$TSS = Y'M_\iota Y = ... = (M_\iota \hat{Y})'(M_\iota \hat{Y}) + e'e = ESS + RSS$$

- We have seen uncentered R^2 in the previous example.
- However, uncentered R^2 is not invariant to changes of unit.
- If $Y + \alpha\iota$, we end up with two R^2s. To see this:

$$R_a^2 = \frac{\|P_X Y\|^2}{\|Y\|^2} = \cos^2\theta \leq 1$$

$$R_b^2 = \frac{\|P_X Y + \alpha\iota\|^2}{\|Y + \alpha\iota\|^2} \neq R_a^2.$$

Centered R-square

- Now cosider the expression of deviation from mean for all variables.

$$
\begin{aligned}
Y &= \iota\beta_1 + X_2\beta_2 + \varepsilon, \quad X = (\iota : X_2) \\
M_\iota Y &= M_\iota X_2\beta_2 + \text{residuals}.
\end{aligned}
$$

- The R^2 using centered variables is defined as

$$
R_c^2 = \frac{\|M_\iota P_X Y\|^2}{\|M_\iota Y\|^2} = 1 - \frac{\|M_X Y\|^2}{\|M_\iota Y\|^2}.
$$

- Remark: Now the centered R_c^2 won't be affected by the addition of a constant to the regressand.
- Remark: If there is no intercept term in the regression, R_c^2 will not make sense in this setting. The reason is that $\sum_{i=1}^n e_i/n = \bar{e} \neq 0$.

Adjusted R-square

- Observe that both R_c^2 and R_u^2 are non-decreasing in the number of regressors.
  i.e., you may want to include the # of regressors without bound.
- You will get the redundancy problem.
- There is a popular modification to R^2.
- Define the adjusted R-square, $\bar{R}^2$:

$$\bar{R}^2 \equiv 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = R_c^2 - \frac{k-1}{n-k}\left(1 - R_c^2\right),$$

- Can you see the trade-off?

Change of units

- Researchers may measure the explanatory variables in different units, such as gram or kilogram
  - sometimes the scale is too large like population: `1,300,000,000`
- Can do the transformation on data matrix by post multiplying nonsingular matrix D with $\dim(D) = k$.
- For instance, pick $D = \text{diag}(d_1, d_2, ..., d_k)$, which rescales all the regressors in the original regression.
- Let's look at the change by $D = \text{diag}(d_1, d_2, ..., d_k)$

$$Y = X\beta + \varepsilon \Rightarrow Y = XD\gamma + \eta$$

- New estimator:

$$\hat{\gamma} = \left(D'X'XD\right)^{-1} D'X'Y = D^{-1}\left(X'X\right)^{-1}\left(D'\right)^{-1} D'X'Y = D^{-1}\hat{\beta}$$

Change of units

Remark

1. *The new parameter estimator changes* ($\hat{\gamma} = D^{-1}\hat{\beta}$)
2. *The fitted values do not change* ($= X\hat{\beta}$)
3. *RSS and* R^2 *are not changed*

Standardized Regression Coefficients

- Sometimes it is helpful to work with scaled explanatory variables and outcome variable that produces dimensionless regression coefficients.
- These dimensionless regression coefficients are called as standardized regression coefficients.
- Unit normal scaling:

$$X_{ik}^* = \frac{X_{ik} - \bar{X}_k}{s_{x_k}}$$

$$Y_i^* = \frac{Y_i - \bar{Y}}{s_y},$$

where $s_{x_k}^2 = \sum_{i=1}^n (X_{ik} - \bar{X}_k)^2/(n-1)$ and $s_y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/(n-1)$.

Statistical Properties

Two theorems below discuss the small-sample properties of $\hat{\beta}$ and $\hat{\sigma}^2$.

Theorem

Given the assumption that $E[\varepsilon|X] = 0$, we can show that $\hat{\beta}$ is an unbiased estimator of β.

Proof.

$$
\begin{aligned}
\hat{\beta} &= (X'X)^{-1} X'Y = \beta + (X'X)^{-1} X'\varepsilon \\
\mathrm{E}[\hat{\beta}|X] &= \beta + \mathrm{E}[(X'X)^{-1} X'\varepsilon|X] \\
&= \beta + (X'X)^{-1} X'\mathrm{E}[\varepsilon|X] = \beta \\
\mathrm{E}[\hat{\beta}] &= \mathrm{E}[\mathrm{E}[\hat{\beta}|X]] = \mathrm{E}[\beta] = \beta.
\end{aligned}
$$

$\square$

Statistical Properties

Remark

In traditional approach, it is assumed that X is non-stochastic or predetermined. However, in the present context it is more reasonable to suppose that X is stochastic.

Remark

Note that condition $E[X'\varepsilon] = 0$ is not enough to ensure the unbiased property of $\hat{\beta}$. The essential assumption is $E[\varepsilon|X] = 0$. A much stronger condition is that (X_i, ε_i) i.i.d., X_i and ε_i are independent $\forall i$ and $E[\varepsilon_i] = 0$

Statistical Properties

Theorem

Given the assumption that $E[\varepsilon|X] = 0$, $E[\varepsilon\varepsilon'|X] = \sigma^2 I_n$, *we can show that* $\hat{\sigma}^2 = e'e/(n-k)$ *is an unbiased estimator of* σ^2.

Proof.

$$
\begin{aligned}
\mathrm{E}[\frac{e'e}{n-k}|X] &= \frac{1}{n-k}\mathrm{E}\left[Y'M_X M_X Y|X\right] = \frac{1}{n-k}\mathrm{E}[\mathrm{tr}\left(Y'M_X Y\right)|X] \\
&= \frac{1}{n-k}\mathrm{E}[\mathrm{tr}\left(\varepsilon' M_X \varepsilon\right)|X] = \frac{1}{n-k}[\mathrm{tr}(M_X \mathrm{E}\left[\varepsilon\varepsilon'|X\right])] \\
&= \frac{\sigma^2}{n-k}[\mathrm{tr}\left(M_X\right)] = \frac{\sigma^2}{n-k}(n-k) = \sigma^2.
\end{aligned}
$$

$\square$

Statistical Properties

- The variance-covariance matrix of $\hat{\beta}$ is

$$\begin{aligned} \text{var}[\hat{\beta}|X] &= \mathsf{E}[(\hat{\beta} - \mathsf{E}[\hat{\beta}])(\hat{\beta} - \mathsf{E}[\hat{\beta}])'|X] \\ &= \mathsf{E}[\left(X'X\right)^{-1} X'\varepsilon\varepsilon'X \left(X'X\right)^{-1} |X] \\ &= \sigma^2 \left(X'X\right)^{-1}. \end{aligned}$$

- Using the idea in Lec2, one could prove the Gauss-Markov Theorem in multiple regression framework.
- Sometimes $\hat{\beta}$ is called minimum variance linear unbiased estimator (MVLUE).
- Of course, other estimators (biased or nonlinear classes) of β may have smaller variance than $\hat{\beta}$.